ATTENTION: this is a very early draft of a paper, or rather a developing project. The final paper will be significantly different. We show it to our colleagues and collaborators to collect feedback and suggestions. On rare occasions it is being shown to people who want to get a first idea of "what it is all about".
PLEASE DO NOT PROPAGATE THIS DOCUMENT.
WE APPRECIATE YOUR CONSIDERATION (A.Osterman)

The Subsystems Approach to Genome Annotation and its Use In the Project to Annotate 1000 Genomes

By Ross Overbeek and many co-authors
(Fellowship for Interpretation of Genomes, and many other institutions)

Abstract

We anticipate that the release of the 1,000th complete genome will occur in the next two to three years. In anticipation of this milestone, the Fellowship for Interpretation of Genomes (FIG) launched the **Project to Annotate 1,000 Genomes** in December 2003. The project is built around a central strategic principle: *the key to improved accuracy in high-throughput annotation technology is to have experts annotate single subsystems over the complete collection of genomes*. For example, an expert in a specific metabolic pathway should annotate the genes that implement that pathway (across all genomes), rather than having an annotation expert attempt to annotate all of the genes in a single genome. Using the first approach, all of the genes implementing the subsystem are analyzed by an expert in that subsystem, while the second approach almost assures that all of the genes will be annotated by an individual with no special expertise.

Once the project committed to the annotation of subsystem across the entire collection of genomes, it became necessary to create an annotation environment in which populated subsystems can be curated and projected to new genomes. The key to development of this technology was to define a portable notion of *populated subsystem*, and then to provide tools for exchanging and curating these objects. The SEED is an annotation environment that supports this model of annotation, but the central capability of maintaining and extending populated subsystems will almost certainly be included in more annotation environments as the utility becomes apparent.

With the publication of this paper, we are making the first release of our growing library of populated subsystems available. It contains xxx subsystems that include genes from yyy organisms.

Introduction

In the 10 years since the first complete bacterial genome was released in 1995, more than 200 complete genomes have been sequenced. There has been an exponential growth in the number of complete genomes sequenced, and based on past growth we anticipate that the 1,000th genome will be sequenced at some point during 2007 (Fig. 1). This rapid release of data requires high-throughput annotation systems that provide a reliable, and accurate, annotation of as many genes as possible.

In response to these challenges the Fellowship for Interpretation of Genomes (FIG) launched the "Project to Annotate a 1,000 Genomes". The Project embodies a specific strategic view of how to approach high-throughput annotation: the effort is organized around individuals who master the details of a specific subsystem and then analyze and annotate the genes that make up that given subsystem over an entire collection of genomes. A subsystems based approach provides many benefits compared to more traditional techniques of genome annotation.

1. The analysis of a single subsystem over a large collection of genomes can produce far more accurate annotations than the more common approach of annotating the genes within a single organism. The usual approach ensures that in most cases the individual annotating a specific gene lacks any specific expertise in the overall physiology related to the role of the gene. Moreover, the usual approach also limits evidence that may support or refute a particular annotation. For example, genome context analysis techniques require that the genes of multiple organisms be considered in parallel. Indeed, as the number of sequenced genomes increases, the value of comparative analysis improves dramatically.

2. The annotation of protein families rather than an organism at a time brings to bear specialized expertise and consequently leads to improvements over initial annotations. Analysis of families offers a significant improvement over analysis of individual genes; however analysis of *sets of related protein families* (i.e., those containing genes that make

up a single biological subsystem) is often more productive than analyzing single families in isolation.

3. It is both more straightforward and less error prone to automatically extend annotations from a set of populated subsystems covering a diverse set of organisms than to extend annotations using the existing automated pipelines.

4. Coverage of a large number of subsystems in diverse organisms is critical to advance other bioinformatics efforts such as metabolic reconstruction, stoichiometric modeling and gene discovery.

Most of the existing annotation tools are not designed to allow cross-genome annotation of gene ensembles that constitute subsystems. An international collaboration has developed a novel annotation system, the SEED.

What Is Meant By the Term "Subsystem"?

By the term subsystem, we refer to a collection of *functional roles* that together implement a specific biological process or structural complex. A subsystem may be thought of as a generalization of the term *pathway*. Thus, just as glycolysis is composed of a set of functional roles (glucokinase, glucose-6-phosphate isomerase, phosphofuctokinase, etc.) a complex like the ribosome or a transport system can be viewed as a collection of functional roles. The genes in each specific organism that includes the subsystem are thought of as implementing those functional roles. In this very general use of the term *subsystem* we make no distinction between metabolic subsystems (i.e., metabolic pathways) and non-metabolic subsystems.

Subsystem definitions can include a collection of functional roles broad enough to cover distinct variants. For example, a subsystem often contains sets of closely related alternative functional roles (e.g., ATP-dependent, ADP-dependent, and pyrophosphate-dependent phosphofuctokinases), or alternative reaction paths that make up the general subsystem. Exactly which roles make up a subsystem is somewhat arbitrary and defined by the user. We rely on

domain experts to choose the precise boundaries of subsystems (i.e., to choose the specific set of functional roles that will be considered together as a set). The SEED annotation system does not limit the set of functional roles that can be defined. As in nature, where a single gene product may contribute to multiple cellular processes, multiple subsystems may have overlapping roles.

By the term *populated subsystem* we refer to a subsystem encapsulated in a matrix in which each column represents a functional role for the subsystem, each row represents a specific genome, and each cell contains those genes from the specific organism that implement the specific functional role. Examples of populated subsystems are given in Fig. 2 and Fig. 3.

Each protein-encoding gene has an associated set of *subsystem connections*, which is simply the set of functional roles that the gene product is considered capable of implementing. In the case of a gene encoding a multifunctional product, we specify the set of subsystem connections as a text string in which the distinct functional roles are separated by slashes (e.g., *Phosphoglycerate kinase / Triosephosphate isomerase).*

A Controlled Vocabulary for Expression of Gene Function

The use of controlled vocabularies (and corresponding ontologies) within genomics is justifiably gaining attention. We believe that the technology that we are developing relates directly to the implementation of controlled vocabularies for expression of gene function. The term "gene function" has come to have several meanings. To understand the position we are adopting, it is useful to distinguish four concepts:

1. A *functional role* is an abstract function such as "Aspartokinase (EC 2.7.2.4)". When we say that a subsystem is a set of functional roles, we mean that it is a set of such abstract functions. We allow the developers of subsystems to specify a single, precise text string to represent each functional role.
2. By the notion of *product name* we refer to a short text string that someone has used to represent the function of the protein encoded by a gene. There are no constraints on the

strings used as product names, and it is common to see the same abstract function denoted by numerous similar expressions.

3. By the term *protein family* we mean a collection of proteins that have been grouped by some curator. The UniProt effort is producing one particularly valuable collection of families. Within that effort, the protein family represents a set of proteins that share a common domain structure. That is, they may actually implement the same or multiple functional roles. As our understanding progresses, we would anticipate being able to break specific families into subfamilies in which each member implements the same set of functional roles.
4. The notation of annotation is often used to refer to an unstructured text string associated with specific genes and/or proteins.

To illustrate our use of these terms, consider the product name "thermostable aspartokinase". It quite possibly is associated with a gene that includes a subsystem connection to the functional role "Aspartokinase (EC 2.7.2.4)", which a curator has included in the subsystem "Lysine_Biosynthesis_DAP_Pathway". The curator may well have attached an annotation describing the specific journal reference justifying the use of the specific product name or subsystem connection.

Product names are often used to include special properties (e.g., "thermostable" or "lysine-sensitive"), and in some cases they simply include clues of function expressed in a short string (e.g., "*similar to death associated protein kinase"*). The subsystem connections, on the other hand, must unambiguously reference a specific functional role that has been included in the definition of a subsystem. It is certainly possible to implement a framework in which each functional role has an associated set of synonyms, and as long as each synonym unambiguously corresponds to a single functional role precision can be maintained. We have elected to avoid the use of synonyms. Each curator of a subsystem is expected to define the set of functional roles corresponding to the subsystem and to establish subsystem connections based on only those predefined functional roles.

Conflict Resolution

The number of populated subsystems grew rapidly during the early phases of The Project, and subsystems were used to encode diverse concepts running the gamut from metabolic pathways, ribosomal proteins, pathogenicity islands, through essential and conserved genes. As we discuss below, the project was based on a distributed effort in which individuals developed their subsystems on numerous machines. Users developed subsystems (and the associated subsystem connections) separately and deposited them in a central clearinghouse. They downloaded whatever subsystems they wished to install locally.

Different users in remote locations began developing overlapping subsystems (in the sense that different subsystems often included the same functional role). Often the precise strings used to represent the functional role were not identical. This need not be viewed as undesirable. For example, one might well think of a situation in which a specific gene

- had a product name of *thermostable aspartokinase,*
- had a subsystem connection to the functional role *Aspartokinase (EC 2.7.2.4)* in the subsystem "Lysine_Biosynthesis_DAP_Pathway", and
- had a subsystem connection to *aspartokinase* in the subsystem "Methionine Biosynthesis".

In such a case, the set of subsystem connections would include connections to the functional roles {*Aspartokinase (EC 2.7.2.4), aspartokinase*}.

Within the Project to Annotate a 1000 Genomes, we decided that this was undesirable. We have sought to maintain a single formulation of each functional role. To support uniform terminology requires that conflicts be detected whenever subsystems are imported and that conflicts be resolved (by renaming functional roles).

To facilitate coordination and communication between end users, to aid with conflict resolution, and to eliminate redundancy, we have developed a multi-author website using Wiki technology. The site (http://www-unix.mcs.anl.gov/SEEDWiki/moin.cgi/SubsystemBulletinBoard) provides an overview of the subsystems that are currently in progress and highlights individual researchers efforts. For a more detailed discussion of each of the subsystems we have also developed a bulletin-board system (powered by vBulletin software). The bulletin board available at (http://www.subsys.info/) has subsystems separated by class, and each subsystem has a discussion arena for the deposition of comments, questions, suggestions, and ideas.

The Project to Annotate 1000 Genomes has collected populated subsystems and integrated them into a single collection (with a consistent formulation of functional roles) which we are now releasing. We wish to emphasize that the existence of an open source collection of the software tools, along with freely available versions of other researchers' work via the clearinghouse, creates a rare opportunity for independent groups to construct and distribute alternative collections.

Subsystems: a Technology Independent of Annotation System

Before describing the system we have constructed and support as a framework for curating subsystems, it is important to emphasize that we are implementing a technology that is independent of any specific annotation system. The overall strategy may be expressed quite concisely:

1. We have defined a simple, portable text representation of a populated subsystem. This allows populated subsystems to be exchanged, archived and updated.
2. We have implemented a *clearinghouse*. Curators *publish* populated subsystems to the clearinghouse, which simply deposits the current version from the local environment to the clearinghouse.
3. Anyone has unrestricted access to the published subsystems. We supply instructions on how to both publish and access the collection.

Once the utility of subsystems technology becomes established, we believe that numerous distinct annotation environments will be developed to support creation and curation of subsystems. It is important to note that we have implemented a strategy that does not require any centralized resource (e.g., a pathway collection) and can straightforwardly be implemented using a variety of software technologies.

Annotation technology needed to support subsystems

In order to support the Project to Annotate 1000 Genomes, we chose to build an annotation environment upon two pre-existing annotation systems. Since the development of subsystems is the central component of the Project, we have implemented a set of tools that have been used to develop this initial collection that we are now releasing.

The institutions participating in the Project have constructed a system supporting initial identification of genes (for prokaryotic genomes), generation of automated assignments of function, creation of populated subsystems, and publishing subsystems to the clearinghouse, and downloading subsystems developed at other sites. The major software components were developed as part of the SEED and the GenDB software efforts. The SEED was developed by an international collaboration led by members of FIG and Argonne National Laboratory. GenDB was developed by a team at the University of Bielefeld, Germany. Both systems were constructed to support annotation of genes and genomes. The teams have integrated these systems into a single execution environment and the resulting software is being made available as open source software (http:xxx). This system offers rudimentary capabilities for defining, curating and exchanging populated subsystems.

The enhancements which need to be added to any existing annotation system to support analysis of subsystems include functionality

1. to encode populated subsystems as objects,
2. to define functional roles and initial subsystems,
3. to connect protein-encoding genes to functional roles in specific subsystems,

4. to publish populated subsystems, and
5. to download subsystems from the clearinghouse.

Once this basic functionality is present, annotation systems can differentiate themselves in many ways, and hopefully the result will be rapidly improving annotation environments which can all be used to create and exchange populated subsystems.

Example Populated Subsystems

The initial release is available from http://theseed.uchicago.edu/FIG. To illustrate the advantages of subsystem annotations over "traditional" annotation systems we describe several subsystems below.

Leucine degradation and HMG-CoA Metabolism

The leucine catabolism pathway and populated subsystem is depicted in Fig. 2. An earlier analysis of this subsystem performed by a group annotating a Brucella melitensis genome {DelVecchio, 2002 #14} was presented in {Overbeek, 2003 #186}.

In humans leucine catabolism is coupled to sterol biosynthesis via a hydroxymethylglutaryl-coenzyme A (HMG-CoA) intermediate. The human subsystem has been well characterized because many individual steps have been implicated in common genetic defects. For example, isovaleric academia, methylcrotonylglycinuria, methylglutaconic aciduria, and 3-hydroxy-3-methylglutaric aciduria are all caused by lesions in this pathway. Moreover, the human enzyme HMG-CoA reductase is a target for statin drugs broadly used in cardiovascular disease therapy because it is a rate-limiting step in sterol biosynthesis.

In contrast, until recently only the early degradative steps had been characterized in bacterial genomes—no genes were directly connected to enzymatic steps beyond isovaleryl-CoA (metabolite II in the panel B of Fig. 2). Attempts to project from known eukaryotic genes based exclusively on homology searches produced ambiguous results because most of the enzymes in this pathway are members of large families of paralogs. Such paralogs usually retain a "general

class" function (such as dehydrogenase or carboxyl transferase) but differ widely in substrate specificity. A combination of functional and genome context analysis, as depicted in the populated subsystem spreadsheet (Fig. 2, panels C and D, respectively) allowed us to gather convincing evidence for the presence of the entire pathway of leucine catabolism in a number of diverse bacteria.

We identified a large conserved gene cluster containing reliable bacterial orthologs of two known human genes committed to this pathway – *yngH* is an ortholog of MCCC2 and *yngG* is an ortholog of HMGCL. This observation enabled the refinement of functional annotations for two additional bacterial genes in the same cluster (*yngJ* is an ortholog of IVD and *yngF* is an ortholog of MCCC1). Although these are weak homologs they could not be accurately characterized without tgaking into account the clustering.

When the initial analysis was performed, no sequence data was available for the methylglutaconyl-CoA hydratase (MGCH) from any species, either bacterial or eukaryotic ((Overbeek, Devine et al. 2003)). Based on the observed clustering, we predicted the conserved bacterial gene (*yngG* in Fig. 2), originally annotated as a member of the enoyl-CoA hydratase family, was the methylglutaconyl-CoA hydratase. This functional prediction was projected from *Bacillus* to the human homolog (AUH). The experimental verification of this human gene was recently provided by two independent publications (L, Loupatty et al. 2002; Ly, Peters et al. 2003). This illustrates the direct impact of chromosomal clustering in prokaryotic genomes on annotation and prediction of eukaryotic genes. More examples of such a cross-taxon projection are provided in {Osterman, 2003 #178}. Another functional inference from the cluster analysis is a connection between leucine catabolism and acetoacetate metabolism (as illustrated in Fig. 2, anel B)., This observation suggests a physiologically relevant extension of this subsystem beyond its traditional boundaries (compare Fig. 2 with KEGG:map00290 at http://www.genome.jp/dbget-bin/show_pathway?map00290)

Chromosomal clustering combined with analysis of protein fusion events revealed two forms of *yngF* (encoding MCCC1): the most common form is a fusion of biotin carboxylase and a C-terminal BCCP domain; and a rarer form in which the biotin carboxylase and the downstream

BCCP gene are separate (as in *B. subtilis*). This observation contributes to interpretation of an evolutionary history of a large and diversified group of proteins involved in biotin-dependent transcarboxylation (for an insightful genomic overview, see {Jordan, 2003 #185})

Panels B and C in Fig. 2 illustrate the analysis of functional variants of a subsystem. Most of the subsystem genes are conserved in those species that have a functional ("nonzero") variant. *E. coli* and *S. aureus* do not have a functional variant and we predict that they are incapable of catabolizing leucine using this pathway. They are marked "0" in the "functional variant" column of the subsystem spreadsheet (Fig. 2, panel C), even though *S. aureus* has genes related to utilization of HMG-CoA in the mevalonate pathway of isoprenoid biosynthesis. A distinction between the functional variants 1–3 is made based on the downstream component of the subsystem, the alternative routes of conversion of acetoactetate to succinate (intermediate V in Fig. 1, panel B). This is either via SCOTA/SCOTB (variant 2; e.g. *Brucella melitensis*) or via AACS (variant 3; e.g. *Geobacter metallireducens* and *Shewanella oneidensis*). Both routes are possible in variant 1, as exemplified by both humans and *B. subtilis*, although clustering on the chromosome suggests that in the latter species an AACS-dependent reaction may be preferred or at least co-regulated with the other components of the subsystem.

Coenzyme A Biosynthesis Subsystem.

Coenzyme A is a universal and essential cofactor in all forms of cellular life acting as a principal acyl carrier in numerous biosynthetic, energy-yielding, and degradative pathways (Begley et al. 2001). Earlier bioinformatics analysis of Coenzyme A biosynthesis revealed a number of interesting variations between species {Gerdes, 2002 #87;Osterman, 2003 #178;Genschel, 2004 #326}. Our current version of this subsystem in SEED (see Fig. 3) contains 14 functional roles (including several alternative forms of a single role, see below) and covers 263 diverse genomes (239 bacteria, 13 archaea and 7 eukaryotes). A five-step pathway from panothenate (vitamin B_5) to Coenzyme A is the universal component of the subsystem conserved in the majority of species. The most variable aspect of this pathway is pantothenate kinase (PANK). All of the three known nonorthologous forms of PANK can be detected in bacterial genomes, and, in some cases, two alternative forms are present in the same organism. One of these forms, PANK3, was

recently identified in *B.subtilis* (gene *coaX*, [patent]) and later confirmed in other bacterial species (such as *H.pylori*, Dr. E. Strauss, personal communication), Although PANK3 appears to be wider represented in the bacterial world than the "classic" PANK, it is incorrectly annotated in most genomes, usually being called a "BVG accessory factor". We believe that this is largely because most of the annotation efforts (with a few exceptions) are disconnected from functional context analysis (as implemented by SEED subsystems). PANK2, characteristic of all eukaryotes, was predicted {Daugherty, 2002 #105} and subsequently verified {Choudhry, 2003 #323} as the only PANK in all Staphylococci. No archaeal gene for PANK has been identified so far, although reasonable candidates may be proposed, such as PAE3407 from *Pyrobaculum aerophilum*. Members of this uncharacterized family of putative GHMP-like kinases are conserved in all archaea and have a tendency to cluster on the chromosome with other genes involved in CoA biosynthesis. For example, in the case of *P. aerophilum*, this gene is next to PAE3409 and PAE3410 coding for 2-dehydropantoate 2-reductase (KPRED in Fig. 3) and 3-methyl-2-oxobutanoate hydroxymethyltransferae (KPHMT in Fig. 3), respectively. This conjecture (also suggested by {Genschel, 2004 #326}) requires experimental verification. Based on a long-range sequence similarity analysis, another protein family conserved in archaea (e.g. PAE1629 of *P. aerophilum*) was proposed to fulfill the role of dephospho-coA kinase (DPCK in Fig. 3) (see also COG0237 at NCBI).

Several examples illustrating major functional variants of the subsystem are outlined in Fig.3. (for a more detailed analysis see Y.Ye et al, 2005 submitted to IMBC).

- A de novo pantothenate biosynthesis, which is present in many bacteria (variant 1 and 2), fungi (variant 2) and archaea (variant 3) is functionally replaced by salvage in higher multicellular eukaryotes and a number bacterial pathogens (variant 4).
- The most radically truncated version of subsystem (variant 6) is observed in all Chlamydiaceae and Rickettsiaceae. These intracellular pathogens, are most likely dependent on the salvage of the last CoA precursor (dephospho-CoA) from the eukaryotic host.
- A sub-set of genes present in a small group of bacterial pathogens (such as *Mycoplasma penetrans* and *Treponema palidum*, variant 7) may be rationalized via a "pantetheine shunt". This hypothetical route assumes salvage of the host's pantetheine, a possible

product of poorly understood CoA catabolism. It is supported by an observed efficient phosphorylation of pantetheine by PANK in vitro (Strauss and Begley 2002. Although this variant appears to be rarely used for CoA de novo biosynthesis, it may be a rather common route of pantetheine recycling.

- A disrupted pattern (missing PANK, PPCS and PPCDC) observed in *Buchnera aphidicola* suggests another interesting possibility, a metabolic exchange between this obligate intracellular endosymbiont and aphid host cells. According to this hypothesis, pantothenate produced but not utilized by *B. aphidicola* may be fed directly into the universal pathway of the host. The latter may "pay back" the endosymbiont by providing phosphopantetheine intermediate required for the last two steps of CoA synthesis in *B. aphidicola*.
- Among open (missing gene) problems within this subsystem, there is no candidate for aspartate decarboxylase (ASPDC) in ~ 45 bacterial and fungal genomes with an otherwise complete set of genes for the de novo synthesis (e.g. variants 2 and 5). Possible interpretations of this observation are: (i) the presence of an alternative non-orthologous and presently unknown form of ASPDC (missing gene) or (ii) the existence of an alternative source of beta-alanine ("missing" pathway). Many aspects of CoA biosynthesis are still unclear in archaea. An identity of holo-ACP phosphodiesterase gene involved in CoA catabolism remains an open question, since a previous assignment of this activity to an *acpD* gene of E.coli was experimentally disproved by {Nakanishi, 2001 #325}

The Impact of Populated Subsystems

The utility of subsystem-based annotations can be split into two broad categories: the primary utility of supporting research in the biological subsystems themselves, and a secondary utility arising from their use in addressing numerous fundamental problems within bioinformatics.

The primary utility of populated subsystems relates to the following:

1. A populated subsystem often supports substantially more accurate assignments of function to genes.
2. Analysis of the populated subsystem allows one to arrive at a precise notion of which forms (i.e., which variants) of the subsystem exist in which organisms.
3. The matrix included in a populated subsystem often makes it vividly clear that a gene implementing a specific functional role is very likely to exist, even though it has not yet been identified. These so-called *missing gene* problems occur with surprising frequency. We include in the supplementary material instances in which conjectures could easily be formulated once the actual presence of a missing gene had been identified.
4. The presence of an extensive set of populated subsystems lays the foundation for an accurate characterization of the metabolic network present in each organism (i.e., forms the basis for constructing an accurate stoichiometric matrix).

The existence of a collection of populated subsystems also has an impact on a number of important topics in bioinformatics:

1. Over and over as we performed our analysis we found that genes that appeared to be missing from a populated subsystem were present in the genome but had not been identified as a gene or ORF. For the functional roles represented in populated subsystems, it becomes possible to directly search for instances of these roles in cases in which there is reason to believe that such a gene must exist. Our collaborators at the University of Bielefeld, Germany, have recalled the genes in over 250 prokaryotic genomes, and the existing body of populated subsystems is used to directly support and annotate many of their newly identified genes. A strategy to support targeted searches for specific genes automatically is being developed.
2. An unresolved issue in microbial gene calling algorithms is correctly identifying the start sites of genes. The most successful attempts have been used sequence alignments of related genes. We believe that use of genes that are both similar and believed to implement the same functional role will lead to substantial improvements over existing estimates. Our collaborators at Middle Tennessee State University are investigating enhanced start-site identification based on subsystems analysis.

3. The search for regulatory sites in upstream regions of related genes relies on accurate ORF calls, annotations, and start-site identifications. With the release of our initial set of populated subsystems, we are making data available to support such analysis. For each populated subsystem, we are providing sequences of upstream regions for each prokaryotic genome. Each sequence contains 300bp of upstream sequence depicting the boundary of the adjacent gene (delimiting the intergenic gap), as well as 100bp of the gene sequence itself.
4. The development of carefully curated protein families has historically been a key goal of bioinformatics for obvious reasons. The limitations of existing formulations relate to ambiguities in function assignment, a problem that is directly addressed by populated subsystems. We have used this initial collection to propose some improvements to existing UniProt annotations, and our analyses are available to UniProt and other database efforts to aid in producing clean, comprehensive collections of protein families.
5. Some of the most successful applications of bioinformatics technology relate to *context analysis* [{Overbeek,Osterman,2003}]. In numerous cases, the clues that led to conjectures of function were based on the fact that related genes tend to cluster on prokaryotic chromosomes, tend to fuse, and tend to co-occur. The populated subsystems offer a framework for establishing the statistical properties needed to effectively exploit these tendencies.

As the Project to Annotate 1000 Genomes progresses, a fairly large set of curated subsystems will emerge. We will move from the current state of initial versions of subsystems still containing numerous errors to a set maintained and checked by experts in each domain. As expert-curated subsystems grow in number, we will move towards the goal of tracing the evolutionary histories of the catalytic domains that make up each subsystem (see Ancient Origin of the Tryptophan Operon and the Dynamics of Evolutionary Change, Xie, Keyhani, Bonner, and Jensen, MMBR, Sept 2003, p 303-342 for a detailed illustration of this style of analysis).

The development of an accurate phylogenetic context for the proteins that make up each subsystem will be lead to tools used to address numerous specific questions. In particular we would argue that an accurate phylogenetic framework will lay the foundations for interpretation

of environmental samples. The degree to which one can infer metabolic capabilities based on phylogenetic neighborhood will become clear only as the details infrastructure emerges.

Finally, we would note that access to the data residing in populated subsystems will soon need to be accessible via web services. This single, simple step will immediately lead to broadening the utility of the subsystems collection.

Specific Accomplishments

The construction of the existing set of populated subsystems has led to numerous conjectures and insights. In several cases, populated subsystems have been completed to the point where they support a detailed review, and these will be published separately. In others, conjectures have been formulated but experimental verification is yet to follow. We have illustrated how such conjectures are formed with the two examples we presented earlier. The reader should note that the supplementary material includes descriptions of over XX open problems with significant conjectures for YY of the cases.

The Release

With the publication of this paper, we are making an initial release of our collection of populated subsystems (which is a subset of those available via the SEED clearinghouse). We are making this subset available in a format that makes the data easily accessible for use in other systems or as raw data accessible by tools yet to be constructed.

The current release of xxx populated subsystems is available without restriction from our ftp site at (http://ftp.TheFIG.info/AnnotatedSubsystems). The precise format of the data is described in detail in the release. We provide as many sequence IDs as possible. For example, a particular sequence we include identifiers from FIG, UniProt, KEGG, and NCBI (including GI number, gene number, UI, or RefSeq ID), as well as identifiers from sequencing laboratories.

The data can easily be examined using a publicly accessible installation of the SEED at the University of Chicago (http://theseed.uchicago.edu/FIG will provide access to the SEED, while http://theseed.uchicago.edu/FIG/subsys.cgi will provide a menu of populated subsystems installed on that system).

The SEED itself is open source software and can be acquired in the form of DVDs with instructions from seed-tech@mcs.anl.gov. It runs on both Mac OS X systems and Linux systems.

Conclusions

Within 2-3 years we will all have access to over a thousand sequenced genomes. This data will grow to become the central resource in modern biology. Annotating this collection is the core challenge of modern bioinformatics. In this paper we describe a new approach to annotation based on idea of subsystems that promises to dramatically improve the quality and utility of annotations. This approach is central to the Project to Annotate 1,000 genomes and has been implemented in a suite of tools for genome annotation. The approach and technology provide one way to involve many domain experts in the genome annotation process.

The technology for developing these subsystems now exists, the technologies for supporting automated addition of new genomes to the collection of populated subsystems is now being developed, and the initial collection is being made available to the research community. Although this initial release contains many errors and omissions, we believe that it will ultimately constitute the foundation upon which future annotation of genomes will be based.

References

1. Schuler, G.D., et al., *Entrez: molecular biology database and retrieval system.* Methods Enzymol, 1996. 266: p. 141-62.

2. Cathy H. Wu, H.H., Leslie Arminski, Jorge Castro-Alvear, Yongxing Chen, Zhang-Zhi Hu, Robert S. Ledley, Kali C. Lewis, Hans-Werner Mewes, Bruce C. Orcutt, Baris E. Suzek, Akira Tsugita, C. R. Vinayaka, Lai-Su L. Yeh, Jian Zhang, and Winona C. Barker, *The Protein Information Resource: an integrated public resource of functional annotation of proteins.* Nucleic Acids Ressearch, 2002. 30: p. 35-37.

3. Ogata, H., et al., *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Res, 1999. 27(1): p. 29-34.

4. Dmitrij Frishman, M.M., Denis Kosykh, Gabi Kastenmu¨ ller, et al., *The PEDANT genome database.* Nucleic Acids Research, 2003, 2003. 31(1): p. 207-211.

5. Bairoch, A. and R. Apweiler, *The SWISS-PROT protein sequence data bank and its new supplement TREMBL.* Nucleic Acids Res, 1996. 24(1): p. 21-5.

6. Etzold, T., and Argos, P., *SRS an indexing and retrieval tool for flat file data libraries.* Comput. Appl. Biosci., 1993. 9: p. 49-57.

7. Etzold, T., Ulyanov, A., and Argos, P., *SRS: Information Retrieval System for Molecular Biology Data Banks.* Methods in Enzymology, 1996. 266: p. 114.

8. Overbeek, R.a.P., M., *Accessing Integrated Genomic Data Using GenoBase: A Tutorial (Part 1).* 1993, Argonne National Laboratory.

9. T. Gaasterland, N.M., R. Overbeek, and E. Selkov. *PUMA: An Integration of Biological Data to Support the Interpretation of Genomes.* in *3rd Albany Conference on Computational Biology "Phylogenetic and Structural Relationships among Proteins".* 1995.

10. Graham, D.E., N. Kyrpides, I. J. Anderson, R. Overbeek, and W. B. Whitman, *Genome of Methanocaldococcus (Methanococcus) jannaschii.* Methods Enzymol, 2001. 330: p. 40-123.

11. Klenk, H.P., Clayton, .A., Tomb, ., White, ., Nelson, .E., et al., *The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus.* Nature, 1997. 390: p. 364-370.

12. Deckert G, W.P., Gaasterland T, Young WG, Lenox AL, Graham DE, Overbeek R, Snead MA, Keller M, Aujay M, Huber R, Feldman RA, Short JM, Olsen GJ, Swanson RV, *The complete genome of the hyperthermophilic bacterium Aquifex aeolicus,* Nature, 1998. 392: p. 353-358.

13. Overbeek, R., et al., *WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction.* Nucleic Acids Res, 2000. 28(1): p. 123-5.

14. Selkov Jr., E., Grechkin, Y., Mikhailova, N., andSelkov, E., *MPW: the Metabolic Pathways Database. Nucleic Acids Res*., 1998. 26: p. 43-45.

15. Overbeek, R., et al., *The ERGO genome analysis and discovery system.* Nucleic Acids Res, 2003. 31(1): p. 164-71.

16. Muhl, G., *Large-Scale Content-Based Publish/Subscribe Systems*. 2002, Darmstadt University of Technology: Darmstadt, Germany.

17. Balakrishnan, I.S.a.R.M.a.D.K.a.M.F.K.a.H. *Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications*. in *ACM SIGCOMM '01 Conference*. 2001. San Diego, California.

18. Enrico Franconi, G.K., Andrei Lopatenko, Luciano Serafini. *A Robust and Computational Characterisation of Peer-to-Peer Database Systems*. in *International Workshop On Databases, Information Systems and Peer-to-Peer Computing*. 2003. Berlin, Germany.

19. Aberer, K., Despotovic, Z. *Managing Trust in a Peer-2-Peer Information System*. in *Tenth International Conference on Information and Knowledge Management (CIKM01)*. 2001.

20. *Peer to Peer Security Subgroup*, Internet2.

21. Stevens, R., Papka, M., Disz, T., *Prototyping the Workspaces of the Future.* IEEE Internet Computing, 2003. 7(4): p. 51-58.

22. Belokosztolszki, A., Eyers, D., Pietzuch, P., Bacon, J., Moody, K. *Role-Based Access Control for Publish/Subscribe Middleware Architectures*. in *2nd International Workshop on Distributed Event-Based Systems (DEBS'03)*. 2003. San Diego, California.

23. Terpstra, W., Behnel, S., Fiege, L., Zeidler, A., and Buchmann, A. *A Peer-to-Peer Approach to Content-Based Publish/Subscribe*. in *Proceedings of the 2nd International Workshop on Distributed Event-Based Systems (DEBS'03)*. 2003. San Diego, CA.

24. Doval, D., O'Mahony, D., *Overlay Networks: A Scalable Alternative for P2P.* IEEE Internet Computing, 2003. 7(4): p. 79-82.

25. Etzold T, V.G. *Using views for retrieving data from extremely heterogeneous databanks.* in *Pac Symp Biocomput.* 1997.

26. Zdobnov EM, L.R., Apweiler R, Etzold T., *The EBI SRS server-new features.* Bioinformatics., 2002. 18(8): p. 1149-50.

27. *The IETF Peer to Peer Research Group Charter*.

28. *PyGlobus*, Lawrence Berkeley National Laboratory.

29. Instituite, A.N.S., *Role Based Access Control*. 2003.

30. *OASIS eXtensible Access Control Markup Language TC*.

31. Rajesh Raman, M.L., Marvin Solomon. *Matchmaking: Distributed Resource Management for High Throughput Computing*. in *HPDC*. 1998.

32. DelVecchio, V.G., et al., *The genome sequence of the facultative intracellular pathogen Brucella melitensis.* Proc. Natl. Acad. Sci. USA, 2002. 99(1): p. 443-8.

33. Paulsen, I.T.e.a., *The Brucella suis genome reveals fundamental similarities between animal and plant pathogens and symbionts. Proc Natl Acad Sci U S A* 99 (20), 2002: p. 13148-13153.

34. Kurnasov, O.V., Polanuyer, B.M., Ananta, S., Sloutsky, R., Tam, A., Gerdes, S.Y.& Osterman, A.L., *Ribosylnicotinamide kinase domain of NadR protein: Identification and implications in NAD biosynthesis.* Bacteriology, 2002. 184: p. 6906-17.

35. Singh, S.K., Kurnasov, O.V., Chen, B., Robinson, H., Grishin, N.V., Osterman, A.L. & Zhang, H., *Crystal structure of Haemophilus influenzae NadR protein: a bifunctional enzyme endowed with NMN adenylyltransferase and ribosylnicotinamide kinase activities. J. Biol. Chem*., 2002. 277: p. 33291-9.

36. Begley, T., Kinsland, C, Mehl, RA, Osterman, A. & Dorrestein P, *The biosynthesis of nicotinamide adenine dinucleotides in bacteria. Vitam Horm*, 2001. 61: p. 103-19.

37. Zhang, H., Zhou, T., Kurnasov, O., Cheek, A., Grishin, N.V. & Osterman, A., *Crystal Structures of E. coli nicotinate mononucleotide adenylyltransferase and its complex with deamido-NAD. Structure*, 2002. 10: p. 69-79.

38. Zhang, X., Kurnasov, O.V., Karthikeyan, S., Grishin, N.V., Osterman, A.L., Zhang, H., *Structural characterization of a human cytosolic NMN/NaMN adenylyltransferase and implication in human NAD biosynthesis. J. Biol. Chem.*, 2003.

39. Zhou, T., Kurnasov, O., Tomchick, D.R., Binns, D.D., Grishin, N.V., Marquez, V.E., Osterman, A. & Zhang, H., *Structure of human NMN/NaMN adenylyltransferase: basis for the dual substrate specificity and activation of anticancer drug tiazofurin. J. Biol. Chem*., 2002. 277: p. 13`148-54.

40. Kurnasov, O., Goral1, V., Colabroy, K., Gerdes S., Anantha S., Osterman, A., Begley T., *NAD Biosynthesis: identification of the tryptophan to quinolinate pathway in bacteria. Chemistry and Biology*: p. Submitted.

41. Kurnasov O., J.L., Polanuyer B., Dorrestein P., Begley T, A. Osterman, *Aerobic Tryptophan Degradation Pathway in Bacteria: Novel Kynurenine Formamidase. FEMS letters*, 2003. In Press.

42. Daugherty, M., Polanuyer, B., Farrell, M., Lykidis, A., de Crécy-Lagard, V. & Osterman, A., *Human coenzyme A biosynthesis: complete pathway reconstitution via comparative genomics. J. Biol. Chem*., 2002. 277: p. 21431-9.

43. Karthikeyan, S., Zhou, Q., Mseeh, F., Grishin, N.V., Osterman, A.L. & Zhang, H., *Ligand-binding Induced Conformational Changes in Riboflavin Kinase: Structural Basis for the Ordered Mechanism. Biochemistry*.

44. Karthikeyan, S., Zhou, Q., Mseeh, F., Grishin, N.V., Osterman, A.L. & Zhang, H., *Crystal structure of human riboflavin kinase reveals a barrel fold and a novel active site arch. Structure*, 2003. 11: p. 265-273.

45. Gerdes, S., Scholle, M., D'Souza, M., Bernal, A., Baev, M., Farrell, M., Kurnasov, O., Daugherty, M., Mseeh, F., Polanuyer, B., Campbell, J., Anantha, S., Shatalin, K., Chowdhury, S., Fonstein, M. & Osterman, A., *From genetic footprinting to antimicrobial drug targets: Examples in cofactor biosynthetic pathways. J. Bact*, 2002. 184: p. 4555-72.

46. Gerdes, S., Scholle, M., Campbell, J., Balazsi, G., Ravasz, E., Daugherty, M., Somera, A.L., Kyrpides, N., Anderson, I., Gelfand, M.S., Bhattacharya, A., Kapatral, V., D'Souza, M., Baev, M., Mseeh, F., Fonstein, M., Overbeek, R., Barabasi, A.-L., Oltvai, Z.N. & Osterman, A., *Experimental determination and system-level analysis of essential genes in E. coli MG1655. J. Bact.,* 2003: p. In Press.

47. Osterman, A., Overbeek, R., *Missing genes in metabolic pathways: a comparative genomics approach.* Curr Opin Chem Biol, 2003. 7(2): p. 238-51.

48. Gerdes, S.Y., et al., *From genetic footprinting to antimicrobial drug targets: examples in cofactor biosynthetic pathways.* J Bacteriol, 2002. 184(16): p. 4555-72.

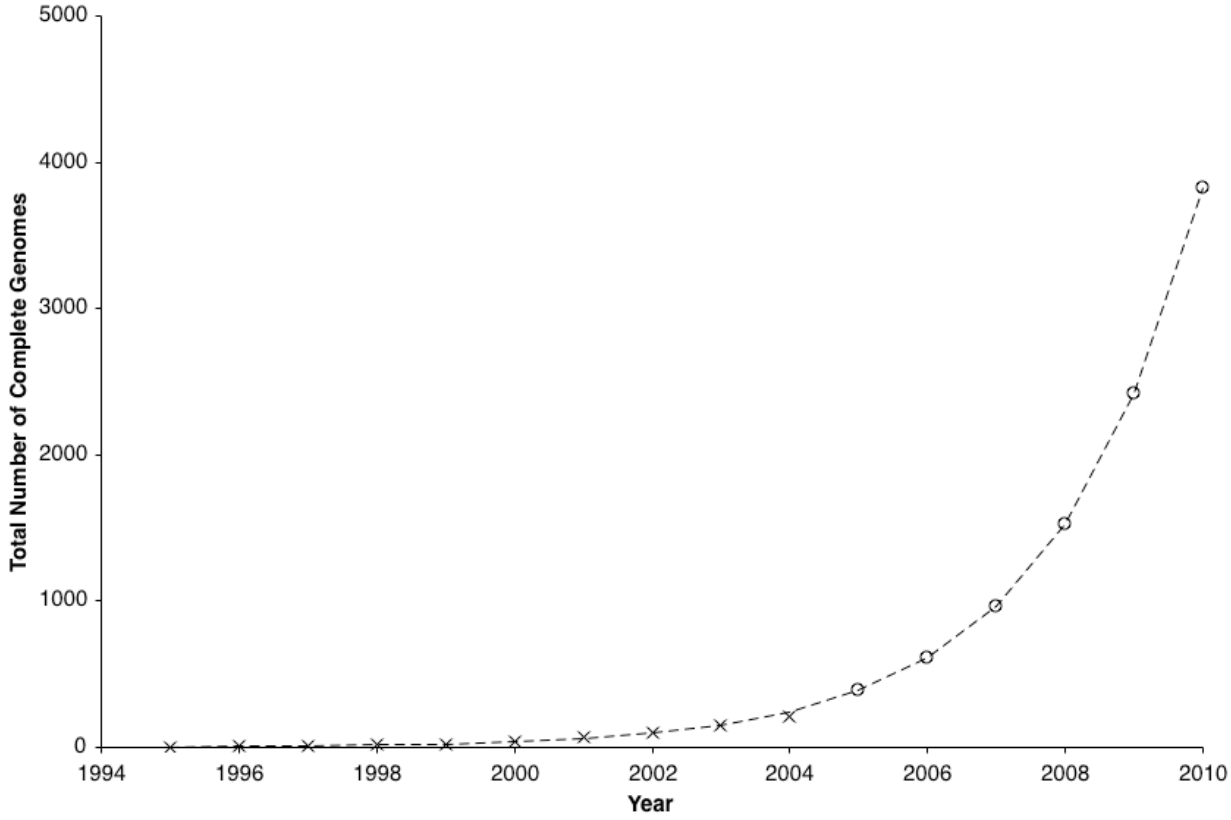
Figure Legends

Fig. 1. Accumulation of complete archeal and bacterial genome sequences at NCBI 1994-2004, and prediction of the release of genomes through 2010. Data from http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi was extracted and plotted by year as shown with the crosses. Data from 2004-2010 is projected by the power law and is represented by open circles. At the current rate of growth, the 1000th complete microbial genome will be released in late 2007 or early 2008.

Fig. 2 Leucine degradation and HMG-CoA Metabolism Subsystem. A) Functional roles in the subsystem. B) Subsystem diagram: key intermediates (circles with roman numerals), connected by enzymes (boxes with abbreviations matching the upper panel) and reactions (arrows). Presence of genes assigned with respective functions is shown for two genomes, *Brucella melitensis* and *Geobacter metallireducens*, using color-coded highlighting as explained in the panel. C) Subsystem spreadsheet presence of genes assigned with functions is shown by gene names for *B. subtilis* or by "+" for all other genomes (modified from a regular SEED display showing all protein IDs). Highlighting by a matching color indicates proximity on the chromosome. D) Clustering on the chromosome of genes involved in the Subsystem demonstrated by alignment of the chromosomal contigs of respective genomes around a signature pathway gene, yngG. Homologous genes are shown by arrows with matching colors and numbers corresponding to functional roles in panel A. *B. subtilis genes* are marked by gene names. Other genes (not conserved within the cluster) are colored gray.

Fig. 3. Coenzyme A Subsystem.

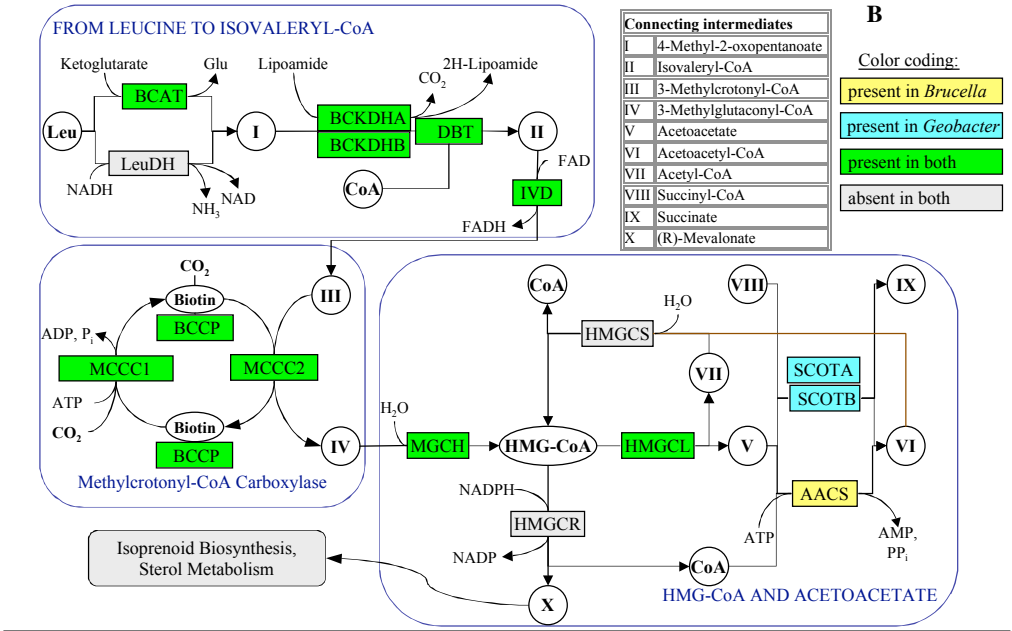
Overbeek *et al.,* Fig. 1.



Overbeek *et al.*, Fig. 2

A

Subsystem: Leucine degradation and HMG-CoA Metabolism		
1	**BCAT**	Branched-chain amino acid aminotransferase (EC 2.6.1.42)
2	**LDH**	Leucine dehydrogenase (EC 1.4.1.9)
3	**BCKDHA**	2-oxoisovalerate dehydrogenase alpha subunit (EC 1.2.4.4)
4	**BCKDHB**	2-oxoisovalerate dehydrogenase alpha subunit (EC 1.2.4.4)
5	**DBT**	Lipoamide acyltransferase of dehydrogenase complex (EC 2.3.1.-)
6	**IVD**	Isovaleryl-CoA dehydrogenase (EC 1.3.99.10)
7	**MCCC1**	Carboxyl transferase chain of methylcrotonyl-CoA carboxylase (EC 6.4.1.4)
8	**MCCC2**	Biotin carboxylase chain of methylcrotonyl-CoA carboxylase
9	**BCCP**	Biotin carboxyl carrier protein of methylcrotonyl-CoA carboxylase
10	**MGCH**	Methylglutaconyl-CoA hydratase (EC 4.2.1.18)
11	**HMGCL**	Hydroxymethylglutaryl-CoA lyase (EC 4.1.3.4)
12	**AACS**	Acetoacetyl-CoA synthetase (EC 6.2.1.16)
13	**SCOTA**	Succinyl-CoA:3-ketoacid-coenzyme A transferase subunit A (EC 2.8.3.5)
14	**SCOTB**	Succinyl-CoA:3-ketoacid-coenzyme A transferase subunit B (EC 2.8.3.5)
15	**HMGCS**	Hydroxymethylglutaryl-CoA synthase (EC 2.3.3.10)
16	**HMGCR**	3-hydroxy-3-methylglutaryl-coenzyme A reductase (EC 1.1.1.34)



Connecting intermediates	
I	4-Methyl-2-oxopentanoate
II	Isovaleryl-CoA
III	3-Methylcrotonyl-CoA
IV	3-Methylglutaconyl-CoA
V	Acetoacetate
VI	Acetoacetyl-CoA
VII	Acetyl-CoA
VIII	Succinyl-CoA
IX	Succinate
X	(R)-Mevalonate

B

Color coding:

present in *Brucella*

present in *Geobacter*

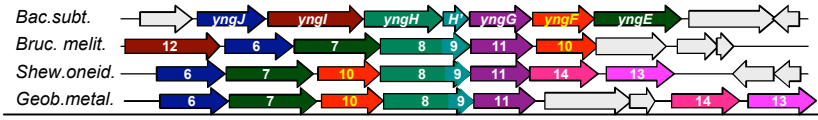
present in both

absent in both

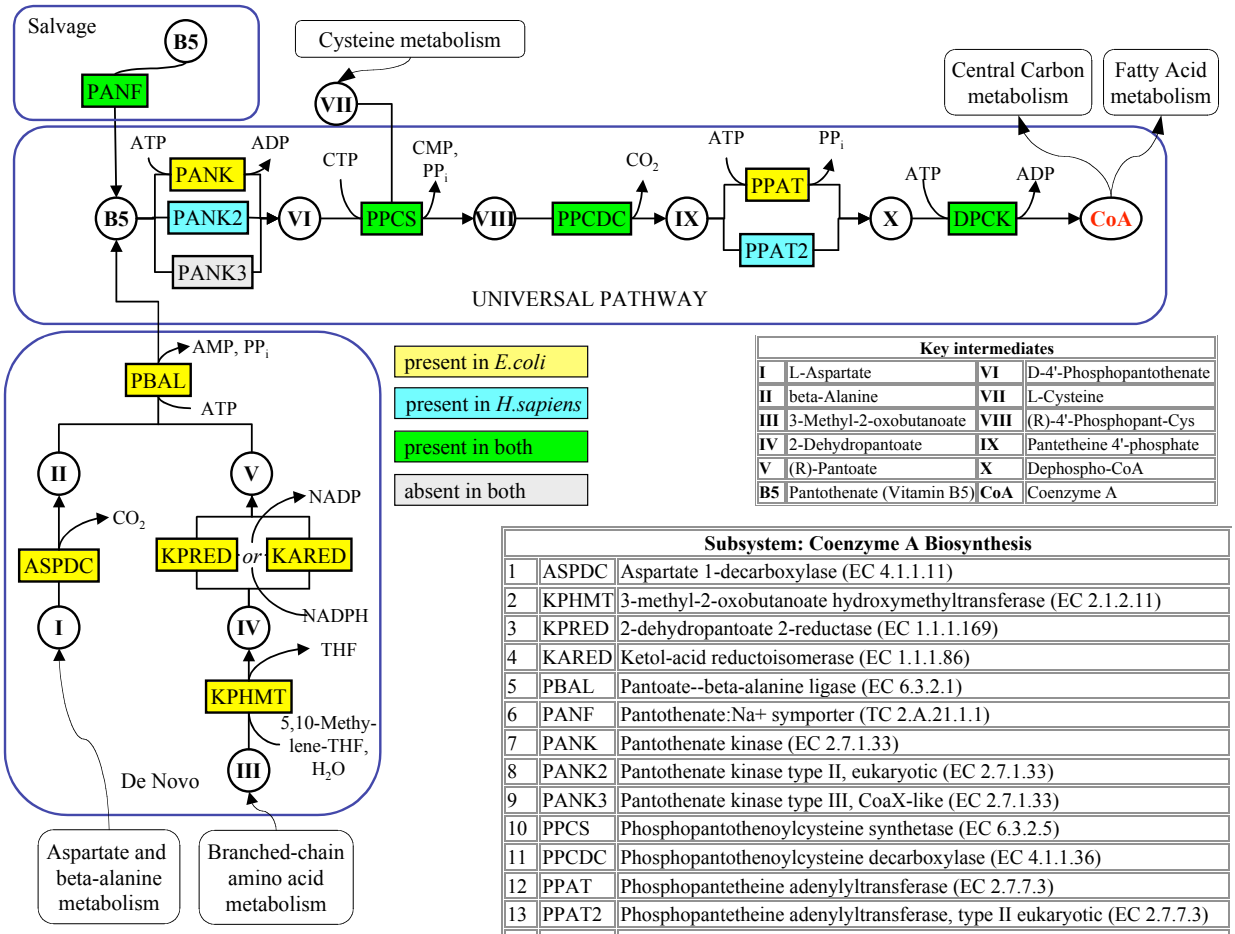
C

Genomes	Funct. variant	From Leucine to Isovaleryl-CoA					To Acetoacetate via HMG-CoA						Acetoacetate			HMG-CoA	
		BCAT	LeuDH	BCKDHA	BCKDHB	DBT	IVD	MCCC1	MCCC2	BCCP	MGCH	HMGCL	AACS	SCOTA	SCOTB	HMGCS	HMGCR
Bac.subtil.	1	*ywaA*	*bcd*	*bkdA*	*bkdB*	*bkdA*	*yngJ*	*yngF*	*yngH*	*yngH'*	*yngF*	*yngG*	*yngI*	*yxjD*	*yxjE*	*pksG*	
Bruc.melit.	2	+		+	+	+	+	+	+	+	+	+	+				
Shew.oneid.	3	+	+	+	+	+	+	+	+	+	+	+		+	+		
Geob.metall.	3	+		+	+	+	+	+	+	+	+	+		+	+		
Hom.sapiens	1	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+
Staph.aureus	0															+	+
E. coli K12	0	+															

D



Overbeek et al., Fig. 3



Key intermediates			
I	L-Aspartate	VI	D-4'-Phosphopantothenate
II	beta-Alanine	VII	L-Cysteine
III	3-Methyl-2-oxobutanoate	VIII	(R)-4'-Phosphopant-Cys
IV	2-Dehydropantoate	IX	Pantetheine 4'-phosphate
V	(R)-Pantoate	X	Dephospho-CoA
B5	Pantothenate (Vitamin B5)	CoA	Coenzyme A

Subsystem: Coenzyme A Biosynthesis		
1	ASPDC	Aspartate 1-decarboxylase (EC 4.1.1.11)
2	KPHMT	3-methyl-2-oxobutanoate hydroxymethyltransferase (EC 2.1.2.11)
3	KPRED	2-dehydropantoate 2-reductase (EC 1.1.1.169)
4	KARED	Ketol-acid reductoisomerase (EC 1.1.1.86)
5	PBAL	Pantoate--beta-alanine ligase (EC 6.3.2.1)
6	PANF	Pantothenate:Na+ symporter (TC 2.A.21.1.1)
7	PANK	Pantothenate kinase (EC 2.7.1.33)
8	PANK2	Pantothenate kinase type II, eukaryotic (EC 2.7.1.33)
9	PANK3	Pantothenate kinase type III, CoaX-like (EC 2.7.1.33)
10	PPCS	Phosphopantothenoylcysteine synthetase (EC 6.3.2.5)
11	PPCDC	Phosphopantothenoylcysteine decarboxylase (EC 4.1.1.36)
12	PPAT	Phosphopantetheine adenylyltransferase (EC 2.7.7.3)
13	PPAT2	Phosphopantetheine adenylyltransferase, type II eukaryotic (EC 2.7.7.3)
14	DPCK	Dephospho-CoA kinase (EC 2.7.1.24)

Examples/ number of genomes	variant	ASPDC	KPHMT	alternatives		PBAL	PANF	alternatives			PPCS	PPCDC	alternatives		DPCK
				KPRED	KARED			PANK	PANK3	PANK2			PPAT	PPAT2	
De novo, complete /100		+	+	+ or/and +		+	±	+or/and + or +			+	+	+		+
Esc.coli K12	1a	*panD*	*panB*	*panE*	*ilvC*	*panC*	*panF*	*coaA*			*coaBC*	*coaBC*	*coaD*		*coaE*
Dein.radiod..	1b	900	2793		1701	1347	659		643		761	761	824		2074
Staph.aur.	1c	2477	2479	2480	1926	2478	1771+			1999	1089	1089	1004		1561
De novo (ASPDC=?)/45		?	+	+ or/and +		+	±	+or/and + or +			+	+	+ or +		+
Shew. onei.	2a	?	810	3466	3948	809		199			3851	3851	4255		389
Geob.metalli.	2b	?	179	2031	2898	178	1470		1816		2004	2004	67		1465
Sach. cerev.	2c	?	377	755	3888	2637				1397	2697	3297+		2314	1076
De novo, archaea/8		?	+	+ or/and +		?	±	?			+	+		+	?
Pyrob.aeroph.	3	?	2402	2401		?		?			831	831		608	?
Salvage of B5/40							+/?	+or/and + or +			+	+	+ or +		+
Strep.pneum	4a				403			741			1109	1110	1781		873
Therm.teng.	4b				15				2203		1410	1410	1388		814
Hom.sap	4c						11370+			10185+	12518	11601		12914	12914+
Truncated pathways/20		±	±	±		±		±					±		+
Buch.aphid	5	?	196		566	195							554		202
Trep.pallid.	6								430				283		296
Chlam.trach.	7														504

Overbeek *et al.,* Supplemental Materials

Some of the examples are briefly outlined in the Supplemental Materials. Those include:

1.Vitamin and Cofactor Metabolism.

NAD and NADP biosynthesis.

FMN and FAD biosynthesis

[… more? Possible: biotin, ubiquinone]

2. Amino acid Metabolism

Lysine biosynthesis (DAP pathway)

Histidine biosynthesis

Histidine degradation

[… more? Possible: Methionine]

3. Carbohydrate Metabolism

UDP-N-Acetylmuramate biosynthesis

N-Acetylglucosamine utilization

[… more?]

4. Fatty acids and lipids

Fatty acid synthase II

Isoprenoid synthesis

[… more? carotenoids]

5. Other contributions by GJO, Sveta et al.